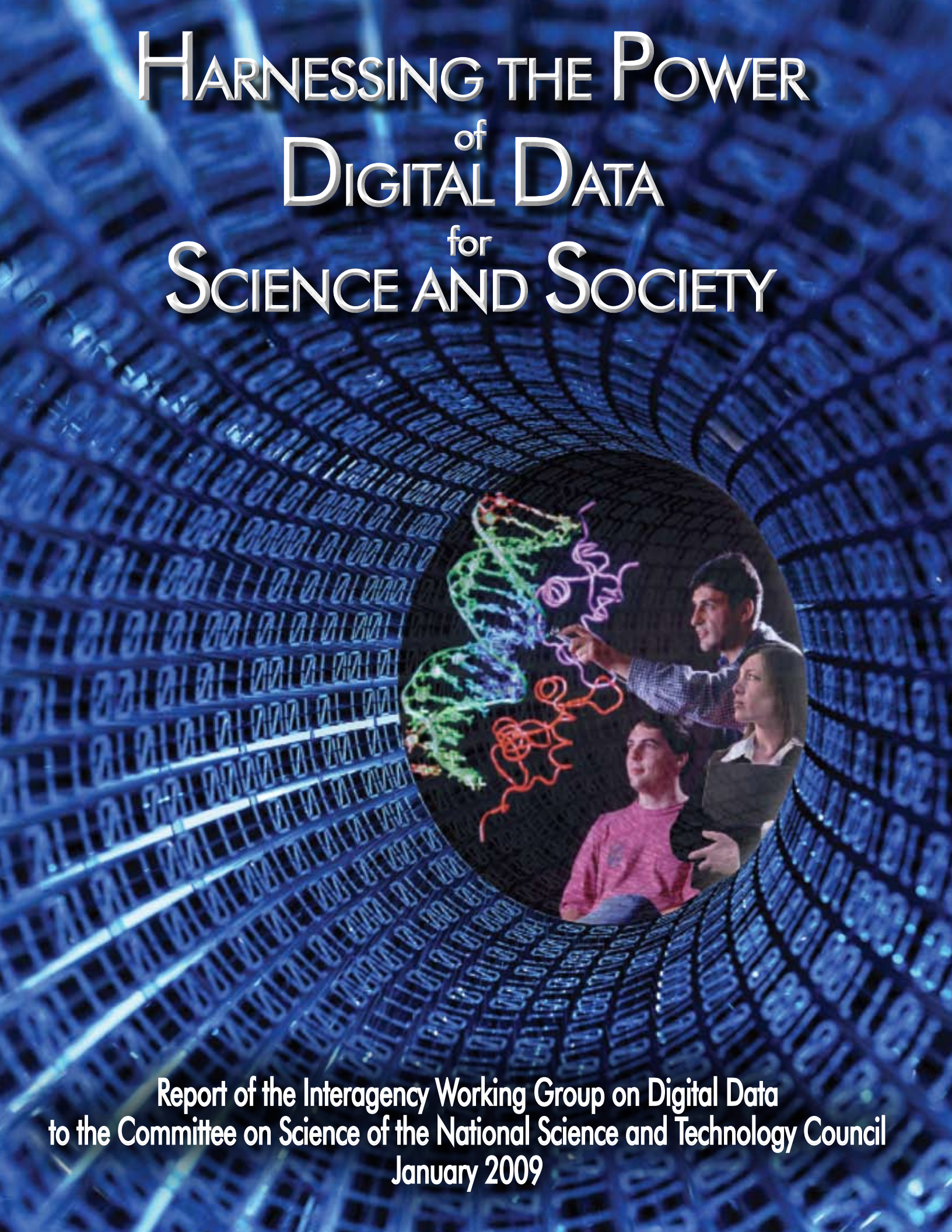
Harnessing the Power of Digital Data for Science and Society



Report of the Interagency Working Group on Digital Data
to the Committee on Science of the National Science and Technology Council
January 2009

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JAN 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Harnessing the Power of Digital Data for Science and Society	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Science and Technology Council,Committee on Science,Washington,DC**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**60**	

Cover Design by Terri S. Lloyd, Information International Associates, Inc.
Cover image courtesy of the Theoretical and Computational Biophysics Group, NIH Resource for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign. Original photo by R. Thompson, modified by Information International Associates, Inc., with the permission of the owner.

HARNESSING THE POWER OF DIGITAL DATA FOR SCIENCE AND SOCIETY

Report of the Interagency Working Group on Digital Data
to the Committee on Science of the National Science and Technology Council

January 2009

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

January 14, 2009

Dear Colleague,

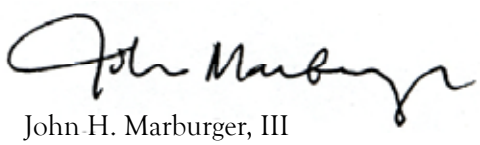
Digital technologies are reshaping the practice of science. Digital imaging, sensors, analytical instrumentation and other technologies are becoming increasingly central to experimental and observational research in all areas of science. Increases in computational capacity and capability drive more powerful modeling, simulation, and analysis to link theory and experimentation and extend the reach of science. Improvements in network capacity and capability continually increase access to information, instrumentation, and colleagues around the globe. Digital data are the common thread linking these powerful trends in science.

Our Nation's continuing leadership in science relies increasingly on effective and reliable access to digital scientific data. Researchers and students who can find and re-use digital data are able to apply them in innovative ways and novel combinations for discovery and understanding. The return on the Nation's investment in generating or acquiring scientific data is multiplied when data are reliably preserved for continuing, creative use. Remote, networked access can lower barriers to participation, allowing citizens in settings throughout the country to benefit from and participate in our Nation's science endeavors.

Responding to the opportunities and needs created by these trends, the National Science and Technology Council's Committee on Science formed the Interagency Working Group on Digital Data. The Group was charged with creating a strategic plan for the Federal government to foster the development of a framework for reliable preservation and effective access to digital scientific data. This report, Harnessing the Power of Digital Data for Science and Society, provides a set of first principles that guide a vision, strategy, tactical goals, and implementation plans for the Federal government, acting as both leader and partner, to work with all sectors of our society to enable reliable and effective digital data preservation and access.

I commend this plan as an important step in addressing the digital data preservation and access needs of our Nation's science and engineering research and education enterprise.

Sincerely,



John H. Marburger, III

Director

Interagency Working Group on Digital Data Participants List

Agency for Healthcare Research and Quality (AHRQ)
Tim Erny

Centers for Disease Control (CDC)
Tim Morris

Department of Commerce (DoC)
National Institute of Standards & Technology (NIST)
Cita Furlani

Department of Commerce (DoC)
National Oceanic and Atmospheric Administration (NOAA)
William Turnbull
Helen Wood

Department of Defense (DoD)
Office of the Director Defense Research & Engineering (ODDR&E)
R. Paul Ryan

Department of Energy (DOE)
George Seweryniak
Walter Warnick

Department of Homeland Security (DHS)
Joseph Kielman

Department of State
Bie Yie Ju Fox

Department of Veterans Affairs
Brenda Cuccherini
Joe Francis
Timothy O'Leary

Food and Drug Administration (FDA)
Randy Levin

Institute of Museum and Library Services
Joyce Ray

Library of Congress (LoC)
Babak Hamidzadeh

National Aeronautics and Space Administration (NASA)
Joe Bredekamp
Martha Maiden

National Archives and Records Administration (NARA)
Robert Chadduck
Kenneth Thibodeau

National Institutes of Health (NIH)
Donald King

National Science Foundation (NSF)
Sylvia Spengler

Networking and Information Technology Research and Development (NITRD)
Robert Bohn
Chris Greer

Office of Science and Technology Policy (OSTP)
Charles Romine

Smithsonian Institution
Martin Elvis
Giuseppina Fabbiano

U.S. Department of Agriculture (USDA/ERS)
Paul Gibson

U.S. Department of Agriculture (USDA/ARS)
Ronnie Green
Kevin Hackett

U.S. Geological Survey (USGS)
Anne Frondorf

IWGDD Executive Secretary
Bonnie Carroll

National Science Foundation (NSF)
Committee on Science Executive Secretary
Marta Cehelsky
Mayra Montrose

Table of Contents

Interagency Working Group on
Digital Data Participants List. iv

Executive Summary 1

Introduction 3

The Current Data Landscape 6

Guiding Principles. 10

Strategic Framework, Recommendations,
and Goals. 13

APPENDIX A

Interagency Working Group on
Digital Data Terms of Reference (Charter) A1

APPENDIX B

Digital Data Life Cycle B1

APPENDIX C

Organizations, Individuals, Roles, Sectors, and Types . . . C1

APPENDIX D

Related Documents D1

Executive Summary

This report provides a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society.

Empowered by an array of new digital technologies, science in the 21st century will be conducted in a fully digital world. In this world, the power of digital information to catalyze progress is limited only by the power of the human mind. Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity. A few bits, well found, can drive a giant leap of creativity. The power of a data set is amplified by ingenuity through applications unimagined by the authors and distant from the original field.

Key characteristics of the current digital data landscape are:

- *the products of science and the starting point for new research are increasingly digital and increasingly "born-digital";*
- *exploding volumes and rising demand for data use are driven by the rapid pace of digital technology innovations;*
- *all sectors of society are stakeholders in digital preservation and access; and*
- *a comprehensive framework for cooperation and coordination to manage the risks to preservation of digital data is missing.*

The following guiding principles were deduced from an analysis of the current digital scientific data landscape. These are based on the expertise of the members of the Interagency Working Group on Digital Data (IWGDD), supplemented by input from outside experts and documentation from major studies of the challenges and opportunities presented by a fully digital world. These guiding principles are:

- *science is global and thrives in the digital dimensions;*
- *digital scientific data are national and global assets;*
- *not all digital scientific data need to be preserved and not all preserved data need to be preserved indefinitely;*
- *communities of practice are an essential feature of the digital landscape;*
- *preservation of digital scientific data is both a government and private sector responsibility and benefits society as a whole;*
- *long-term preservation, access, and interoperability require management of the full data life cycle; and*
- *dynamic strategies are required.*

The strategic framework, recommendations, and goals presented in this report are founded on these guiding principles.

VISION AND STRATEGY

We envision a digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination can be appropriately, reliably, and readily managed. This will enhance the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society.

We set out the following strategy to achieve this vision:

> *Create a comprehensive framework of transparent, evolvable, extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data. Such a framework will serve as a driving force for American leadership in science and in a competitive, global information society.*

RECOMMENDATIONS AND SUPPORTING GOALS

To pursue this strategy, we recommend that:

- *a National Science and Technology Council (NSTC) Subcommittee for digital scientific data preservation, access, and interoperability be created;*
- *appropriate departments and agencies lay the foundations for agency digital scientific data policy and make the policy publicly available; and*
- *agencies promote a data management planning process for projects that generate preservation data.*

Implemented together, these recommendations can reshape the digital scientific data landscape. Through the strength of the NSTC environment, we can pursue goals requiring broad cooperation and coordination while enabling agencies to pursue their missions and empower their respective communities of practice. The goals targeted by these recommendations are:

- *to be both leader and partner;*
- *to maximize digital data access and utility;*
- *to implement rational, cost-efficient planning and management processes;*
- *to empower the current generation while preparing the next;*
- *to support global capability; and*
- *to enable communities of practice.*

Key elements to ensure that these recommendations work together for maximum impact include the following:

- *Subcommittee responsibilities should include topics requiring broad coordination, such as extended national and international coordination; education and workforce development; interoperability; data systems implementation and deployment; and data assurance, quality, discovery, and dissemination.*
- *In laying appropriate policy foundations, agencies should consider all components of a comprehensive agency data policy, such as preservation and access guidelines; assignment of responsibilities; information about specialized data policies; provisions for cooperation, coordination and partnerships; and means for updates and revisions.*
- *The components of data management plans should identify the types of data and their expected impact; specify relevant standards; and outline provisions for protection, access, and continuing preservation.*

Introduction

A REVOLUTION IN SCIENCE

> *"What is at stake is nothing less than the ways in which astronomy will be done in the era of information abundance."*[1]

The fabric of science is changing, driven by a revolution in digital technologies. These include (1) digital imaging devices for astronomy, (2) microarrays and high-throughput DNA sequencers in genomics, (3) wireless sensor arrays and satellites in geosciences, and (4) powerful computational modeling in meteorology. These technologies generate massive data sets that fuel progress. Technologies for high-speed, high-capacity networked connectivity have changed the nature of collaboration and have also expanded opportunities to participate in science through instant access to rich information resources around the world. While these digital technologies are the engine of this revolution, digital data[2] are the fuel.

All elements of the pillars of science – observation, experiment, theory, and modeling – are transformed by the continuous cycle of generation, access, and use of an ever-increasing range and volume of digital data. Experiments and observations can be better designed if a rich set of supporting information is easily accessible. A framework of data can provide a strong foundation on which expansive theory can be developed and refined. Data initiate, drive, and produce dynamic modeling and simulation approaches.

Integrative approaches combine the concepts and tools of many disciplines to take on some of the most important and difficult questions in science. These approaches require the ability to find and use data from many fields and applications. Progress on questions such as (1) the basis for human consciousness and cognition, (2) the nature of dark matter in the universe, and (3) the identification of energy sources that can replace fossil fuels require insights from various disciplines into data of many different types and sources. Global scale science that can meet today's global challenges requires the ability to share and use a distributed array of sources for a wide diversity of information. For example, the workings of the Earth's atmosphere, climate, and interior, and the interplay between economics, culture, politics, and behavior in a global human society, present challenges that require data gathered worldwide. The scale of resources needed for 21st century science often requires global investments, such as the array of instruments needed to explore our universe or a high-energy collider capable of revealing the nature of matter. These resources generate powerful data sets that drive scientific progress around the world.

US National Virtual Observatory

The New Astronomy

All astronomers observe the same sky, but with different techniques, from the ground and from space, each showing different facets of the universe. The result is a plurality of disciplines (e.g., radio, optical or X-ray astronomy and computational theory), all producing large volumes of digital data. The opportunities for new discoveries are greatest in the comparison and combination of data from different parts of the spectrum, from different telescopes and archives.

Astronomers worldwide have recognized this opportunity and have begun a network of collaborations to establish the infrastructure for digital data interoperability. The National Virtual Observatory (NVO) is a partnership of US institutions including universities, observatories, NASA- and NSF-funded centers, and federal agencies including the Smithsonian Institution. The NVO also collaborates with the private sector (e.g., Google and Microsoft), to develop interactive visual portals to the sky. The NVO is a founding member of the worldwide International Virtual Observatory Alliance (IVOA).

Source: NVO: http://www.us-vo.org/; IVOA: http://www.ivoa.net/

1 Towards the National Virtual Observatory: A Report Prepared by the National Virtual Observatory Science Definition Team. See http://www.astro.caltech.edu/~george/sdt/sdt-final.pdf.

2 For purposes of this document, digital data are defined as any information that can be stored digitally and accessed electronically, with a focus specifically on data used by the federal government to address national needs or derived from research and development funded by the federal government.



The LHC: One of the world's most complex data systems

The $3.6 billion Large Hadron Collider (LHC) will sample and record the results of up to 600 million proton collisions per second, producing roughly 15 petabytes (15 million gigabytes) of data annually in search of new fundamental particles. To allow thousands of scientists from around the globe to collaborate on the analysis of these data over the next 15 years (the estimated lifetime of the LHC), tens of thousands of computers located around the world are being harnessed in a distributed computing network called the Grid. Within the Grid, described as the most powerful supercomputer system in the world, the avalanche of data will be analyzed, shared, re-purposed and combined in innovative new ways designed to reveal the secrets of the fundamental properties of matter.

LHC source: public.web.cern.ch/public/en/LHC/LHC-en.html
Source: public.web.cern.ch/Public/en/LHC/LHC-en.html

THE DIGITAL DIMENSION

The digital dimension consists of network connectivity that can lower conventional barriers to participation and interaction of time and place; computational capacity and capability to expand the possible and extend the conceivable; and information discovery, integration, and analysis capabilities to drive innovation. The emergence and continuing evolution of this powerful new dimension is reshaping science, just as it is recasting business, government, education, and many other aspects of human activity worldwide. To lead in the emerging global digital information society, the nation must fully embrace the digital dimension – expanding access, extending capabilities, and building on the potential of this exciting new environment.

The power of digital information to catalyze progress is limited only by the power of the human mind. Data are not consumed by the ideas and innovations they spark, but are an endless fuel for creativity. A small bit of information, well found, can drive a giant leap of creativity. The power of a data set can be amplified by ingenuity through applications unimagined by the authors and distant from the original field. Re-use and re-purposing of digital scientific data have dramatic benefits. First, they provide the basis for doing science at new levels. The reach of a scientist is extended by access to greater inputs than could be gathered by an individual working alone. The goal can be larger and more complex if the products of many different technologies and approaches can be brought to bear. The perspective is exponentially broadened by multiple points of view.

> *"The widespread availability of digital content creates opportunities for new forms of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data. We call this 'cyberscholarship.'"*[3]

Second, preservation to enable re-use and re-purposing ensures maximum return on our nation's investment in science. Effective re-purposing requires interoperability[4] – the ability to combine diverse data, tools, systems, and archives smoothly and simply. By providing for interoperability, genome sequence and protein structure information can be used in innovative combinations to design new drugs to cure and prevent disease and to improve the quality of life. As another example, weather and climate data can be integrated to predict the outbreak of an epidemic.

Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity.

The ability to use data over unlimited time periods and for unlimited purposes creates greater value for science and society.

Third, remote networked access to robust digital information resources changes the participation equation. A student at a tribal college with internet access to a comprehensive

3 The Future of Scholarly Communication: Building the Infrastructure for Cyberinfrastructure. Report of the April 17, 2007 workshop sponsored by the National Science Foundation (NSF) and the Joint Information Systems Committee (JISC) of the United Kingdom.

4 Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged (IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries). The components of interoperability include data, metadata, codes, interfaces, platforms, environments, and networks. Achieving interoperability requires coordination among people, disciplines, and institutions.



With the acquisition of the human genome sequence and the advent of powerful new DNA sequencing technologies and analytical methods, it is increasingly possible to identify variations in human DNA that underlie particular diseases, conditions, or therapeutic responses. The National Center for Biotechnology Information (NCBI) has developed the database of Genotype and Phenotype (dbGaP) to preserve and distribute the results of studies employing these powerful new capabilities. The database represents the combined power of many different types of studies and analyses. As a result, clinicians and scientists from many fields can share their results and work together to investigate the interaction of genotype and phenotype, revealing new links between DNA sequence and a variety of diseases, from breast cancer to diabetes, blood pressure abnormalities, and age-related eye defects.

Source: ncbi.nlm.nih.gov/dbgap

set of digital information resources can contribute according to the quality of ideas.

Fourth, access to digital information supports discovery-based learning, engaging students in the excitement of science. For example, a regional online project allows students recording birds visiting their schoolyards to discover shifts in migratory patterns that are driven by changes in land use. Access also supports innovative research into both new strategies for education and the basis for cognition and learning. Researchers comparing learning patterns across regions or in different settings can uncover some of the influences of culture and context on learning.

Finally, preserving the digital scientific products of our time will ensure that future generations can benefit from our efforts and can better understand our time and place in history.

INTERAGENCY WORKING GROUP ON DIGITAL DATA

Remote networked access to robust digital information resources changes the participation equation.

In December 2006, the National Science and Technology Council of the Committee on Science established the Interagency Working Group on Digital Data (IWGDD; see Appendix A for Terms of Reference). Nearly 30 agencies, offices, and councils were named as members or participants, reflecting the broad range of interests in digital scientific data. The purpose of the IWGDD is to "develop and promote the implementation of a strategic plan for the federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering." This report presents the findings and recommendations of the IWGDD.

The Current Data Landscape



NOAA's DART™ Tsunami Monitoring Buoys

As part of the U.S. National Tsunami Hazard Mitigation Program (NTHMP), the National Oceanic and Atmospheric Administration (NOAA) has developed and placed Deep-ocean Assessment and Reporting of Tsunamis (DART™) stations in regions with a history of generating destructive tsunamis to ensure early detection of tsunamis and to support real-time warnings. Currently DART™ stations are deployed and active in the Pacific, Atlantic and Indian Oceans, the Caribbean Sea, and the Gulf of Mexico.

The tsunami-related data archive has grown from five gigabytes to over 1,700 gigabytes, with standards-compliant metadata available online to support the modeling, mapping, and assessment activities required to minimize the effect of tsunamis.

Source: http://nctr.pmel.noaa.gov/Dart/dart_home.html

An analysis of the current landscape for digital scientific data preservation and access was undertaken through a review of relevant reports and other publications (see Appendix D), agency data policy and strategy documents, and examples of extant digital preservation activities. Highlights of that analysis are presented below.

DIGITAL DATA NEEDS

The conduct of science and engineering is changing and evolving. This is due, in large part, to the expansion of networked cyberinfrastructure and to new techniques and technologies that enable observations of unprecedented quality, detail and scope. Today's science employs revolutionary sensor systems and involves massive, accessible databases, digital libraries, unique visualization environments, and complex computational models.[5]

The use of digital technologies, including computation for increasingly complex models and simulations, vast sensor arrays, powerful imaging equipment and detectors, and networked access, interaction, and dissemination tools, has transformed the scientific landscape. Data that are "born-digital" — available only in digital form and preserved only electronically — are increasingly becoming the primary output of science and the starting point for new research. The rate at which these digital data are produced is increasing each year, yielding massive and exponentially growing data flows in what has been described as a "data deluge."[6]

> *In 2006, the amount of digital information created, captured, and replicated [worldwide] was 1,288 x 10^{18} bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes. This is about 3 million times the information in all the books ever written.*[7]

In principle, a digital data deluge can result in rapid progress in science through wider access and the ability to use sophisticated computational and analytical methods and technologies. In practice, the current landscape lacks a comprehensive framework for reliable digital preservation, access, and interoperability, so data are at risk.

RISK FACTORS

Factors contributing to deterioration or loss of digital data include decay of the storage media; dependence on outmoded formats or systems (hardware and/or software); and errors introduced in reading, writing, or transmission. Additionally, data may be "orphaned" — put at risk of being discarded because the "owner" is no longer identifiable or available. Strategies for mitigating these risks include management planning for data stewardship, controlled redundancy, managed migration to new technologies, and error checking schemes. These promising strategies are limited by two factors. First, many current practices do not scale to the massive volumes and decades-long timelines of many long-term preservation organizations.

5 Investing in America's Future: National Science Foundation Strategic Plan, FY2006-2011.

6 Hey, A. J. G. and Trefethen, A. E. The Data Deluge: An e-Science Perspective. 2003. Berman, Fox, and Hey, Editors. Published in Grid Computing. Making the Global Infrastructure a Reality, pp. 809-824, Wiley and Sons. 2004.

7 The Expanding Digital Universe, IDC White Paper sponsored by EMC Corporation. March 2007.

"It is the contention of the 100 Year Archive Task Force that migration as a discrete long-term preservation methodology is broken in the data center. Today's migration practices do not scale cost-effectively…."[8]

Second, many of these strategies rely on close coordination and cooperation among diverse preservation organizations, but a comprehensive framework is needed to enable coordination and cooperation.

LEGAL AND POLICY LANDSCAPE

The U.S. legal and policy landscape promotes access to digital scientific data produced in the federal and federally funded realms. The elements of this landscape that are most relevant to this document are as follows:

- **The Paperwork Reduction Act** *(44 USC 35) has as one of its key purposes to "ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the federal government."*
- **The Office of Management and Budget (OMB) Circular A-130** *specifies that "The open and efficient exchange of scientific and technical government information … fosters excellence in scientific research and effective use of federal research and development funds."*
- **The 1991 Supreme Court ruling in Feist Publications, Inc. v. Rural Telephone Service Co.** *(499 U.S. 340) establishes that "facts do not owe their origin to an act of authorship, they are not original, and thus are not copyrightable."*
- **Copyright law** *(17 USC 105) provides that "Copyright protection under this title is not available for any work of the United States Government."*
- **The Freedom of Information Act** *(FOIA; 5 USC 552) provides for public access to the records of the federal government.*

This legal and policy landscape produces a climate of equitable access while protecting appropriate intellectual property rights. This provides a dynamic, healthy environment for basic and applied research, enabling the United States to continue as a leader in discovery and innovation in the information age. It also drives a robust commercial information sector. The FY2000 federal investment in public sector information was estimated at $14.9B.[9] The commercial information sector that relies on this investment generated estimated annual sales of $641B, employing 3.2 million people.

ENTITIES IN DIGITAL PRESERVATION AND ACCESS

Many different types of organizations, institutions, groups, and partnerships (referred to below as "entities") are active in the current digital preservation landscape. These include agencies, centers, departments, institutes, libraries, museums, research projects, etc. Over 50 entities across these categories were examined, and the results are outlined in Appendix C. Each entity examined was characterized by:

- *Type (e.g., data center, library, archive, museum)*
- *Roles (e.g., data production, analysis, publication, training)*
- *Sector (e.g., government, research, education)*
- *Expert participants (e.g., librarian, archivist, IT specialist)*

Some of the conclusions emerging from this analysis are described in the following sections.

8 100 Year Archive Requirements Survey, Storage Networking Industry Association. January 2007.

9 Commercial Exploitation of Europe's Public Sector Information: Final Report for the European Commission Directorate General for the Information Society, Pira International. October 2000.

DATA LIFE CYCLE

Most entities currently fulfill multiple roles in the data life cycle, and most roles are being fulfilled by several types of entities. An example is that of data analysis and processing. While this role has traditionally been associated with computational centers, this capability is being implemented in non-traditional settings such as libraries, archives, and museums. This trend toward generalization and away from specialization in the provision of data life cycle functions has important implications. For example, many traditional preservation institutions now operate or require direct access to leading-edge computational facilities, equipment, and expertise, creating new organizational, operational, and financial challenges. Additional implications of this trend toward generalization are discussed in the following sections.

PARTICIPATION BY ALL SECTORS

Nearly all types of preservation entities exist in all sectors – government, education, research institutions, non-profit, commercial, and international. Many entities arise from collaborations across sectors at regional, national, and international levels. An example of a cross-sector partnership is the agreement among the National Archives and Records Administration (NARA), the National Science Foundation (NSF), and the San Diego Supercomputer Center (SDSC) for innovation in preservation of some of the nation's most valuable digital research collections.[10] The Global Biodiversity Information Facility (GBIF) is a partnership of over 70 countries and international organizations providing global access to the world's primary data on biodiversity.[11] This breadth of participation and collaboration provides a potential foundation for sustainability analogous to that provided by diversity in ecosystems sustainability. A digital preservation framework with a diversity of organization types and missions, resources, funding streams, and capabilities is more resilient to changes in short-term trends and to individual failures.

NEW INFORMATION DISCIPLINES

Some new specializations in data tools, infrastructures, sciences, and management are emerging as a result of increased communication and cross-fertilization across the information disciplines that support data preservation. Examples include:

- *Digital Curators: experts knowledgeable of and with responsibility for the content of digital collection(s);*
- *Digital Archivists: experts competent to appraise, acquire, authenticate, preserve, and provide access to records in digital form; and*
- *Data Scientists: information and computer scientists, database and software engineers and programmers, disciplinary experts, expert annotators, and others who are crucial to the successful management of a digital data collection.*[12]

New educational programs and curricula to provide the necessary skill sets and knowledge are beginning to emerge. Viable career paths for some of these areas remain to be developed.

INFORMATION COMMUNITIES[13]

The trend towards generalized data life cycle function does not extend to generalized content. Most preservation entities are closely allied with a particular scientific or topical domain: a community of practice. To increase interoperability within a given science realm, "information communities" are emerging

10 www.archives.gov/press/press-releases/2006/nr06-119.html.
11 www.gbif.org/GBIF_org/participation.
12 Long Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, report of the National Science Board. September 2005.
13 For example, the National Center for Bioinformatics (NCBI) is an information community which draws together genomics scientists, information technologists, evolutionary biologists, and chemists and other communities of practice around a common set of information resources.

that span multiple communities of practice. These communities are working to set their own data standards, establish their own infrastructures, etc. Preservation entities play an important role in this process, providing relevant expertise and experience, as well as a means for implementing and enforcing standards. Significant opportunities exist to promote interoperability and to avoid data "silos" or "stovepipes" which are inaccessible to those outside the immediate community of interest. Instead, interactions among domain-specific entities are encouraged, along with establishment of preservation organizations to span and integrate multiple domains.

PERSONAL DIGITAL COLLECTIONS

With increasing digital access, an individual may have a "personal digital collection" that is specific to and associated with that person. This personal collection may contain data generated by the owner and those drawn transparently from other sources as needed for the analysis at hand. This has two important implications. First, this mode of data use depends on research and development to create powerful new interoperability, data tracking and provenance, attribution, and validation capabilities. Second, ties between users and individual preservation entities may be loosened, threatening models for economic sustainability that depend on these ties. Diversified sustainability models are needed to accommodate this emerging use pattern.



Digitizing Corrosion Information

The Department of Defense is engaged in an ongoing battle with corrosion, which affects most equipment, facilities, vehicles and weapons systems. Making it easier to access results of decades-old corrosion research and technology development could aid in addressing the problem and go a long way toward reducing corrosion-related expenses. To this end, the Advanced Materials, Manufacturing, and Testing Information Analysis Center (AMMTIAC), in a partnership with the Defense Technical Information Center (DTIC), is improving the availability of high-value corrosion research documents from its massive collection of reports with the use of digitization from print and microfiche. The digitized information will be prepared for long-term access and preservation. This endeavor to provide access to full-text resources will, in turn, facilitate the use of sci-tech information associated with corrosion.

NON-DIGITAL COLLECTIONS

"Museums and libraries have leveraged the availability of the Internet to present their resources and services to a broader audience and offered an additional mode of access to them, while traditional in-person visits continue to increase."[14]

Many valuable collections of physical artifacts (documents, books, specimens, etc.) exist in libraries, archives, museums, and other collections throughout the world. Legacy collections of microfiche, audio tapes, film, and other media are housed in repositories, warehouses, and storage facilities around the globe. Digital access to information about these artifacts, or to digital representations of the objects, can greatly enhance awareness and use, increasing the impact of these collections. Strategies, methods, and technologies to create metadata for cataloguing and search/discovery in the digital preservation realm can also inform the non-digital realm. Advantages in digitizing an object include expanded access, enhanced ability to search across collections, and mitigation against catastrophic loss or slow deterioration of the original artifact. Disadvantages can include increased fragility of the digital version and higher costs in some instances for digital versus physical preservation. Decisions about digitization of collections should be based on an evaluation of these advantages and disadvantages, assessed through the combined efforts of digital preservationists and content curators.

14 InterConnections: The IMLS National Study on the Use of Libraries, Museums and the Internet. February 2008.

Guiding Principles

The following guiding principles were deduced from an evaluation of the current digital scientific data landscape. They are based on the expertise and experience of the IWG members supplemented by input from outside experts and documentation from major studies of the challenges and opportunities presented by a fully digital world. The strategic framework, recommendations, and goals presented in this report emerge from these guiding principles.

1. SCIENCE IS GLOBAL AND THRIVES IN THE DIGITAL DIMENSION

The emergence of a powerful new digital dimension brings capabilities for connectivity across oceans and continents, remote access to unprecedented computational power, and the potential to find and use information distributed worldwide. The result is a global landscape in which (1) science can thrive as barriers to collaboration of time and distance are lowered, and (2) limits to the scale, scope, and nature of questions that can be addressed are pushed back by an increasingly capable cyber infrastructure.

2. DIGITAL SCIENTIFIC DATA ARE NATIONAL AND GLOBAL ASSETS

The ability to achieve innovation in a competitive global information society hinges on the capability to swiftly and reliably find, understand, share, and apply complex information from widely distributed sources for discovery, progress, and productivity. Limits on information access translate into limits on all other aspects of competitiveness. Thus, digital information preservation and access capabilities are critical to the progress of individuals, nations, science, and society.

3. NOT ALL DIGITAL SCIENTIFIC DATA NEED TO BE PRESERVED, AND NOT ALL PRESERVED DATA NEED TO BE PRESERVED INDEFINITELY

It is estimated that the amount of digital information produced worldwide each year now exceeds the global digital storage capacity.[15]

Decisions about what to preserve are inevitable. The criteria for such decisions differ among differing data types and contexts. Some data can be reproduced at lower costs than preservation (some outputs of computer models and simulations are examples) and, therefore, may not be a high priority for preservation. Other data cannot be reproduced at any cost (continuous, long-term environmental measurements are examples) and may merit higher priority for preservation. Still other data initially preserved may be superseded by new work and become candidates for disposal. Thus, deliberate decisions about preservation should take place on a continuing basis throughout the full data life cycle. Stakeholders in this decision-making process include: (1) preservation organizations, which must factor their mission, costs, and funding structures into decisions; (2) the scientific community (including communities of practice), which can consider the value to science; (3) data authors, who are most familiar with the detailed context; (4) archival scientists, who bring both an intellectual framework and experience to assessing preservation value; (5) data users, who employ the data in creative and innovative ways; and (6) entities such as associations, federations, and governments, which can take a broad, long-term view.

4. COMMUNITIES OF PRACTICE ARE AN ESSENTIAL FEATURE OF THE DIGITAL LANDSCAPE

Science is conducted in a dynamic, evolving landscape of communities of practice organized around disciplines, methodologies, model systems, project types, research topics, technologies, theories, etc. These communities facilitate scientific progress and can provide a coherent voice for their constituents, enhancing communication and cooperation and enabling processes for quality control, standards development, and validation. These

15 The Expanding Digital Universe, IDC White Paper sponsored by EMC. March 2007.



Barcode of Life

The Barcode of Life Initiative is an international effort to develop reliable and authoritative means for the global identification of biological species. Barcoding uses a short DNA sequence within an organism's genome as the equivalent of a barcode on a supermarket product to determine the species origin of a biological sample. Adoption of a standard format for barcode data allows a sample in a museum or collected in the field to be instantly linked to related information resources worldwide; to be tied in to relevant tissue, parasite, and other collections globally; and to reference DNA databases in the United States, Japan, and Europe. The result is the ability to conduct biodiversity, species migration and invasion, and population genetics studies that are more powerful because they can be reliably compared to and informed by other projects worldwide.

Source: http://barcoding.si.edu/

capabilities are crucial for data preservation and access in communicating the needs and expectations of a community of users, providing expert input on the scientific context for data (including input to decisions about what to preserve and what to discard), promoting good data management practices, and contributing to the development of effective data standards. Thus, data preservation policies and strategies must encourage and enable communities of practice both because of their important role in science and because of the capabilities and perspectives they bring to the preservation process. "One-size-fits-all" policies must be avoided to allow for strategies and designing mechanisms for interoperability that support communities of practice.

5. PRESERVATION OF DIGITAL SCIENTIFIC DATA IS BOTH A GOVERNMENT AND PRIVATE SECTOR RESPONSIBILITY, AND BENEFITS SOCIETY AS A WHOLE

A large number and wide variety of entities, organizations, and communities — each with their own assumptions, culture, expertise, objectives, policies, and resources — are involved in the creation and preservation for access of scientific digital data. Responsibilities for data stewardship are distributed across many diverse entities that, in turn, engage with different institutions, disciplines, and interdisciplinary domains. Responsibility for data stewardship should remain with the distributed collections and repositories that have a vested interest in their community's data. A framework of government/private sector partnerships (analogous to the air transportation or monetary systems) is required to link these distributed responsibilities into an effective system for digital preservation and access.

6. LONG-TERM PRESERVATION, ACCESS, AND INTEROPERABILITY REQUIRE MANAGEMENT OF THE FULL DATA LIFE CYCLE

The full data life cycle includes creation, ingestion or acquisition, documentation, organization, migration, protection, access, and disposition (see Appendix B for a description of the data life cycle) and has two important features. First, the cycle is dynamic rather than static and includes ongoing processes of curation, disposition, and use. Many processes, such as data analyses involving transformation or recombination, are catalytic, continuously increasing the volume of data for preservation and access. Second, the steps in the cycle are not independent. Feasibility, costs, and limitations for each step are strongly dependent on actions taken at other steps. For example, inadequate documentation at an early stage can prevent later use; failure to migrate to new technologies can leave data inaccessible. Effective management of each step and coordination across steps in the life cycle are required to ensure that data are reliably preserved and can be accessed and used efficiently.

Responsibilities for data stewardship are distributed across many diverse entities.

7. DYNAMIC STRATEGIES ARE REQUIRED

"Today, no media, hardware or software exists that can reasonably assure long-term accessibility to digital assets."[16]

The transition from physical to digital information technologies requires several fundamental changes in preservation strategies. First, preservation must be active rather than passive, as data in digital systems are more fragile and the media more transient than in traditional paper- or microfilm-based systems. Second, digital technologies advance continuously, often rendering older technologies unsupported and inaccessible while producing new opportunities for creative exploitation of data. Finally, as remote digital access reduces the need for distributed physical copies, the reduction in systemic redundancy increases the risk of loss. This risk is often managed through redundancy that is actively planned and implemented. In this landscape, recommendations for static solutions are of only transient value. Thus, we focus in this report on processes for actively managing current preservation solutions while continuously anticipating and implementing new methods, technologies, and strategies without endangering preservation and access.

Preservation must be active rather than passive, as data in digital systems are more fragile.

16 The Digital Dilemma, Science and Technology Council of the Academy of Motion Picture Arts and Sciences. 2007.

Strategic Framework, Recommendations, and Goals

The rapid pace of development and deployment of digital technologies are characteristic features of science in the digital dimension. These technologies include digital sensor arrays, powerful imaging technologies, adaptive computing, and increasingly more complex and capable computational modeling and simulation approaches. As a result of these technologies, the volume of digital scientific data is increasing at an exponential rate. This unprecedented growth in digital information presents an equally unprecedented opportunity for progress in all areas of science and engineering research and education if, and only if, the information can be preserved, accessed, understood, and applied. This recommended strategic framework responds to this opportunity with a plan to overcome limits and maximize the scientific information potential of the digital dimension, creating new opportunities and progress for all.

Unprecedented growth in digital information presents an unprecedented opportunity for progress.

VISION

Our strategic vision is a digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination can be appropriately, reliably, and readily managed, thereby enhancing the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society.

STRATEGY

We set out the following strategy to achieve our strategic vision:

Create a comprehensive framework of transparent, evolvable, and extensible policies and management and organizational structures that provide reliable, effective access to the full spectrum of public digital scientific data. Such a framework will serve as a driving force for American leadership in science and in a competitive, global information society.

The framework we envision will allow digital scientific data to be readily discovered, evaluated, and used in creative and complex combinations by specialists and non-specialists alike and will ensure that data are properly protected and reliably preserved. This framework is based on principles for continuous, effective management of new technologies and methods identification and adoption without endangering reliable preservation and access. The essential elements of our strategy are defined as follows:

- *The proposed **"policy, management, and organizational framework"** comprises: (1) an NSTC Subcommittee for digital scientific data preservation and access; (2) the development of agency and organizational data management policies, and (3) data life cycle management planning for relevant projects and activities.*
- ***"Reliable, effective access"** refers to strategies and systems that: (1) provide for reliable, long-term, cost-effective preservation and access at appropriate quality; (2) ensure high-confidence protection of privacy, confidentiality, security, and property rights; (3) enable transparent search and discovery capabilities across a wide range*



A National Map for the 21st Century

The U.S. Geological Survey (USGS) is working with federal, state, and local agencies across the country to create a seamless digital base map for the nation. The goal of The National Map (http://nationalmap.gov) is to become the nation's source for trusted, consistent, integrated, and current topographic information available online for a broad range of uses. By integrating data from many federal, state, and local sources on an ongoing basis, the currency and accuracy of the map are enhanced, making it effective for use in a wide variety of applications such as environmental science and land management, natural hazards and emergency response, and resource planning and decision making.

Source: nationalmap.gov

of resources and data types; (4) include appropriate metadata[17] and documentation to allow data to be understood and effectively re-used or re-purposed; and (5) provide for effective interoperability across repositories, tools, resources, services, and data types and formats.

- ***"Digital scientific data"** refers to born-digital and digitized data produced by, in the custody of, or controlled by federal agencies, or as a result of research funded by those agencies, that are appropriate for use or repurposing for scientific or technical research and educational applications when used under conditions of proper protection and authorization and in accordance with all applicable legal and regulatory requirements. It refers to the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses.*
- ***"American leadership"** among nations worldwide will only be achieved by mobilizing the capabilities of all sectors of our greater society, including government at all levels, industry, foundations, academia, education, and individuals in using, supporting, and evolving the digital scientific data universe.*
- ***"Global information society"** recognizes that science and technology co-exist in a world where technology diminishes geographic, temporal, social, and national barriers to discovery, access, and use of data.*

American leadership will only be achieved by mobilizing the capabilities of all sectors of our society.

This strategy is designed to unite the capabilities and leverage the resources of the federal agencies and organizations in their scientific data activities, thereby enabling the federal government to serve as both leader and partner to all sectors of our society in realizing the full potential of the digital dimension to enable discovery, innovation, and progress.

RECOMMENDATIONS: OUTLINE AND SUPPORTING GOALS

We make three recommendations pursuant to this strategy. These are presented at an outline level below, along with a discussion of the goals that support the recommendations. The final section of this report provides a more detailed discussion of the recommendations. The recommendations are intended to create synergy by combining action items for agencies to work with their communities of practice in pursuit of their respective missions. The recommendations also allow for a forum for cooperation and coordination across government, academic, commercial, and international sectors.

17 Metadata are data about data. They include a formal description of the data, as well as information on how to acquire the data, and information for using the data, such as accuracy, security, and rights. Metadata provide the scientific, technical, contextual, representational, provenance, and other information necessary to enable creative re-use and re-purposing.

RECOMMENDATION 1: WE RECOMMEND THE CREATION OF A NATIONAL SCIENCE AND TECHNOLOGY COUNCIL (NSTC) SUBCOMMITTEE FOR DIGITAL SCIENTIFIC DATA PRESERVATION, ACCESS, AND INTEROPERABILITY.

The NSTC, a Cabinet-level council, is the principal means within the executive branch to coordinate science and technology policy across the federal research and development enterprise. The NSTC's Committee on Science, with its charter to improve the coordination of federal efforts in science, is well positioned to pursue frameworks for cooperation across the federal government that enhance digital scientific data preservation and access. We recommend the creation of a Subcommittee under the Committee on Science to provide the sustained focus and expertise needed to ensure continued leadership in this area. The proposed Subcommittee will provide a mechanism for federal departments and agencies to (1) identify and articulate shared goals for scientific data preservation, access, and interoperability; (2) coordinate planning, implementation, and assessment of their data preservation and access activities; (3) achieve cost-effectiveness by exploiting shared solutions to meet mission requirements and federal standards; (4) provide a means for interaction, collaboration, and coordination with sectors outside the federal arena, including internationally; and (5) coordinate with relevant inter-agency and inter-governmental efforts.

RECOMMENDATION 2: WE RECOMMEND THAT APPROPRIATE DEPARTMENTS AND AGENCIES LAY THE FOUNDATIONS FOR AGENCY DIGITAL SCIENTIFIC DATA POLICY AND MAKE THE POLICY PUBLICLY AVAILABLE.

The appropriate departments and agencies are those who, either directly or through support to others, generate, collect, or steward digital data relevant to science and technology research and education. Data policies should be developed from a foundation of solid understanding and strong consensus on the needs, goals, and best approaches for digital preservation and access both within an agency and across the communities it serves. Currently, agencies are at varying stages in laying the necessary foundations and can benefit from sharing experiences and insights through the forum of the proposed NSTC Subcommittee. With an appropriate foundation in place, agency data policies that address scientific data preservation and access can be developed with community input and in coordination with other departments and agencies. The goals of the agency data policy should be to maximize appropriate information access and utility and to provide for rational, cost-efficient data life cycle management. Agency data policies should be publicly available and should guide and inform the development and implementation of data management plans in individual projects and activities.

RECOMMENDATION 3: WE RECOMMEND THAT AGENCIES PROMOTE A DATA MANAGEMENT PLANNING PROCESS FOR PROJECTS THAT GENERATE PRESERVATION DATA.

In particular, agencies could consider requiring data management plans for projects that will generate preservation data.[18] Advance planning for data preservation and access can ensure that appropriate, cost-effective strategies are identified, and the digital products of research can be made widely available to catalyze progress. Data management plans should provide for the full digital data life cycle and should describe, as applicable, the types of digital data to be produced; the standards to be used; provisions and conditions for access; requirements for protection of appropriate privacy, confidentiality, security, or intellectual property rights; and provisions for long-term preservation (including means for continuously assessing what to keep and for how long).

These recommendations are designed to combine agency actions with interagency and multi-sector cooperation and coordination to pursue the following six goals, which were based on the findings of the IWGDD during its deliberations.

18 "Preservation data" are defined herein as those digital scientific data (either created in digital form or digitized) for which the benefits of preservation are likely to exceed the costs (including the costs of ongoing curation, protection, dissemination, quality control and validation, and migration to new formats and technologies). Inherent in this definition is the need to conduct effective cost/benefit analyses to enable rational decisions about preservation.



The Protein Data Bank Expedites Drug Development

The World-Wide Protein Data Bank (wwPDB) provides a single, authoritative source of information about the structure of biological molecules. Currently, the wwPDB contains the structures of almost 50,000 proteins discovered by researchers worldwide and shared openly with the global community. Many of these proteins, including enzymes, hormones, and receptors, are important drug targets. Because of the quality and quantity of data available via the wwPDB, it is possible to visualize and analyze these molecules to allow engineered drug design. Many of the HIV protease inhibitors used in the cocktail for HIV treatment were developed using this approach. In similar ways, open access to this rich information resource catalyzes progress in many fields and applications.

Source: www.wwpdb.org

GOAL 1: BE BOTH LEADER AND PARTNER

Findings: federal agencies are unique in (1) their responsibilities for gathering data for science; (2) their role in funding scientific research and education; and (3) their ability to make long-term investments, with long-term payoffs, in the interests of society at large. These unique characteristics mean that the federal government must take a leadership role both in providing for preservation and access to digital scientific data and in illuminating the path forward so that others may follow.

It must also be recognized that the digital dimension belongs to all sectors of society. Government at the federal, state, and local levels; industry; academia; foundations; international organizations; and individuals are all participants in the digital dimension and have important interests in and capabilities for digital information preservation and access. Therefore, the federal government has a responsibility to act as a reliable and transparent partner and as a coordinating entity, enabling all sectors to work together in enhancing the information capabilities of the digital dimension.

The digital data challenge cannot be met by the federal government or any one sector acting alone.

The continuing exponential increase in the amount of digital scientific information and the ever-expanding needs and expectations of users exceed both the resources and the mission scope of the federal agencies. The digital data challenge cannot be met by the federal government or any one sector acting alone. The government must act to stimulate and facilitate investments by all sectors of society in meeting the full scope and scale of the scientific data challenge.

To be an effective leader and partner, the federal government must (1) be responsible in meeting respective agency and organizational needs for digital preservation and access; (2) respect and encourage the interests and capabilities of stakeholders in all sectors; (3) be innovative, creating exemplary resources and capabilities to demonstrate feasibility, and establish and disseminate best practices for use in other sectors; (4) provide a coherent mechanism for interaction with other sectors; and (5) promote communication and facilitate partnering among all sectors. Implemented together, the recommendations address all of these responsibilities.

GOAL 2: MAXIMIZE DIGITAL DATA ACCESS AND UTILITY

Findings: Enhanced capabilities for finding, using, and integrating information accelerate the pace of discovery and innovation. Advanced information capabilities and better access to digital data will make America more competitive in a digital world. Thus, a critical requirement for American competitiveness is to establish and continuously improve a robust and pervasive information infrastructure to maximize access to digital scientific data.

Scientific information in an accessible and interoperable digital environment has the characteristics of a public good. The information is not destroyed and its value is not diminished upon use. On the contrary, digital access has a catalytic effect, multiplying the value of information through repeated use by a wide variety of users in a



GEOSS and IEOS
A System of Systems

U.S. Integrated Earth Observation System (IEOS): A Contribution to the Global Earth Observation System of Systems (GEOSS)

Earth observations are the data collected about the Earth's land, atmosphere, oceans, biosphere, and near-space environment. These data are collected by means of instruments that sense or measure the physical, chemical and/or biological properties of the Earth. These data provide critical information to assess climate change and its impacts; ensure healthy air quality; manage ocean, water, mineral and other natural resources; monitor land cover and land use change; measure agricultural productivity and trends; and reduce disaster losses.

The Strategic Plan for the U.S. Integrated Earth Observation System directly supports the efforts of more than 70 countries who are working together to achieve a GEOSS -- interconnecting a diverse and growing array of instruments and systems for monitoring and forecasting changes in the global environment.

Source: http://usgeo.gov/docs/EOCStrategic_Plan.pdf

diversity of settings and applications. This requires effective coordination, extensive interoperability, and innovative tools and services across the full spectrum of digital preservation and access resources.

Digital access has a catalytic effect, multiplying the value of information.

The proposed NSTC Subcommittee is intended to take the lead for the federal government, working in local to global contexts with all sectors of society, to develop mechanisms for maximizing access and utility for digital scientific data. Examples of such mechanisms include: (1) continued improvement in interoperability across all layers (from software to hardware to networks and resources); (2) integration of data from various sources and across projects and disciplines; (3) comprehensive, global, and transparent search, query, and retrieval capabilities; (4) development, continuing evolution, broad adoption, and regular use of appropriate, community-based, cost-effective standards designed to allow efficient information use in innovative ways and in complex combinations; (5) encouragement of digital preservation programs explicitly aimed at facilitating sustained access; (6) promotion of ready access to appropriate documentation and metadata; and (7) reliable protection of security, privacy, confidentiality, and intellectual property rights in complex data environments.

GOAL 3: IMPLEMENT RATIONAL, COST-EFFICIENT PLANNING AND MANAGEMENT PROCESSES

Findings: The total volume of digital data and the rates at which data are being created globally are increasing rapidly. Mobilizing these data in the service of scientific progress without incurring overwhelming costs or risking loss requires robust planning and management processes. These processes must be designed to optimize current resources at all levels, to exploit economies of scale and shared, cost-effective solutions, to anticipate new loads and demands, and to evaluate opportunities and challenges posed by rapidly changing technologies.

The NSTC Subcommittee, working with the appropriate departments and agencies, is well positioned to gather and share across sectors information related to costs and best practices for preservation, protection, dissemination, curation, and migration. This will promote a culture of awareness and capacity for data life cycle management to ensure usable, efficient, cost-effective solutions to data preservation and access.

The process of developing and implementing an agency data policy would be facilitated by the designated agency representative to the Subcommittee. The designee could support the development and maintenance of the agency data policy, ensure that the policy supports the agency mission, provide for appropriate access and preservation of the digital scientific assets, and coordinate with other agencies, sectors, and international partners to further national interests and capabilities. This position requires experience in science, research, and education, and in the full scientific digital data life cycle (see Appendix B).



Earth Observing System Data & Information System (EOSDIS)

The Earth Observing System Data and Information System (EOSDIS) manages and distributes more than 2,700 types of data products and associated services for use in interdisciplinary studies of the Earth system through its eleven data centers. These data centers process, archive, document, and distribute data from NASA's past and current Earth system science research satellites, field programs and aircraft platforms, currently supporting the daily ingest of over 2 terabytes (TB) of satellite instrument data. Over 4.9 petabytes (PB) are archived. In 2007 alone, over 100 million products were distributed to over 165,000 unique users, and approximately 3 million science, government, industry, education and policy-maker users accessed EOSDIS.

The data held at the EOSDIS data centers are interoperable with data from Earth observation communities around the world using a component called the EOS ClearingHOuse (ECHO).

Source: http://outreach.eos.nasa.gov/about.html

GOAL 4: EMPOWER THE CURRENT GENERATION WHILE PREPARING THE NEXT

Findings: To extend the benefits of our strategic vision to all, the education and training to use and manage the current data infrastructure and to develop the future data infrastructure must be widely accessible. If appropriately designed and implemented, the data infrastructure itself can be a robust resource for meeting these education and training needs.

Remaining globally competitive in developing the data capabilities of the future requires both ensuring that future generations of scientists and technologists are capable of operating in the fast-moving world of network and information technologies and providing for the decades-long horizons of digital preservation and access. Assembling an appropriate new cohort of computer and information scientists, cyberinfrastructure and digital technologies experts, digital library and archival scientists, social and behavioral scientists, and others with the requisite skills and expertise to meet this dual challenge can only be done through the combined efforts of the government, education, research, and technology sectors. Key to this effort will be increasing the number of graduates in critical areas such as computer and information sciences and mathematics.

It is crucial that education and training activities be integral to all of the federal science data investments. Facilitating the diffusion of the skills and knowledge necessary to benefit from the digital dimension is essential to achieving our strategic vision and must be integral to all federal science data activities. The NSTC Subcommittee can play a critical role in promoting coordination of education and training among federal departments and agencies and in partnerships with the education, research, and technology sectors. This activity could include the development of joint programs for research and development in the design, implementation, assessment, and evaluation of educational programs.

The nation needs to identify and promote the emergence of new disciplines and specialists expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data. Many disciplines are seeing the emergence of a new type of data science and management expert, accomplished in the computer, information, and data sciences arenas and in another domain science. These individuals are key to the current and future success of the scientific enterprise. However, these individuals often receive little recognition for their contributions and have limited career paths. Critical challenges in achieving our strategic vision include providing an effective pipeline of data professionals to ensure that the needs and opportunities of the future can be met and providing these professionals with appropriate rewards and recognition.

The NSTC Subcommittee will also have an essential role in promoting data science and management as a career path with appropriate recognition and rewards structures. The federal government can be both leader and partner in this arena, using its own programs as models for success and supporting innovative and effective approaches in other sectors. A key goal is to encourage and enable the best and brightest to commit to careers in all aspects of data science to meet the growing needs of our digital society and economy. Further, specialists in



The Case for Biodiversity Data Interoperability

Invading alien species in the United States cause significant environmental damage, with losses adding up to almost $120 billion per year[1]. Cholera bacteria and toxic dinoflagellates have been discovered in ballast water of cargo ships. Yellow fever vectors have spread to new continents in imported tires. Hardwood trees in American cities are being killed by Asian beetles introduced in wooden packing crates. A coordinated, global approach is necessary to detect, understand, and manage the large-scale movement of species. While many electronic databases provide invasive species information, they are not yet fully interoperable. The ability to combine data from a variety of sources is needed to predict and manage invasion threats by interpreting an invasive species' ability to spread into particular regions, calculating its rate of dispersal, and predicting its future range. A global information system that enables interoperability across a diversity of digital resources will require cooperative action at national and international levels.[2]

[1]David Pimentel, Rodolfo Zuniga and Doug Morrison. "Integrating Ecology and Economics in Control Bioinvasions. Ecological Economics," Volume 52, Issue 3, 15 February 2005, Pages 273-288

[2]Excerpted from Anthony Ricciardi, William W. M. Steiner, Richard N. Mack, and Daniel Simberloff, "Toward a Global Information System of Invasive Species," BioScience, Vol 50, No3, March 2000, pp 239-244.

information disciplines (e.g., digital curation and preservation and library and archival sciences) should be given incentives to obtain additional education and training to enable their effective participation in the digital dimension.

Agencies should identify the skills and expertise needed to effectively manage their data resources. The NSTC Subcommittee can be a source of lessons learned and information sharing among the agencies in this regard. Budget planning and cost analyses conducted by the departments and agencies for their data preservation and access activities should consider the costs of education and training programs, including assessment and evaluation, designed to enhance access and utility for their digital resources.

GOAL 5: SUPPORT GLOBAL CAPABILITY

Findings: The digital dimension is global. Science, like many aspects of our global knowledge society, is not limited by national boundaries. Continued U.S. leadership in science will require robust access to information resources, as well as opportunities for collaboration around the world. The American digital preservation and access framework must be effectively international – functionally integrated and closely coordinated with counterparts around the globe.

The digital dimension is global.

These global characteristics and needs will require U.S. investment in (1) shared, international data resources, (2) transparent linkage of U.S. systems and resources to their global counterparts, (3) development, evolution, and integration of appropriate standards, formats, conventions, and other means to provide for interoperability across international boundaries, and (4) efforts to harmonize appropriate legal, regulatory, and policy frameworks to reduce barriers to cooperation, collaboration, and the pursuit of shared goals.

Partnering outside the U.S. requires an accessible point of contact, transparent policy frameworks, and coherence and coordination among federal agencies. The proposed NSTC Subcommittee is well positioned to provide these capabilities and to promote coordination of U.S. data activities with those of our international counterparts.

Where appropriate, data policies developed by departments and agencies should explicitly address plans for achieving global capability. Such policies and plans should identify relevant international stakeholders, processes for standards development and implementation, strategies for enhancing cooperation and coordination to achieve enhanced data access and utility, and mechanisms to identify opportunities for cost savings through economies of scale and sharing of resources within the context of a competitive global economy.



Digital Data Importance to Social and Behavioral Sciences

The study of powerful large-scale trends such as economic development, urbanization, expanding migration, population aging, and mass education by social, behavioral, and other scientists requires access to global-scale micro-data – data about individuals, households, and families collected by census offices around the world. The Integrated Public Use Microdata Series (IPUMS) provides researchers and educators with interoperable access to data from more than 111 censuses in 35 countries representing more than 260,000,000 person records. This powerful digital collection meets critical research needs while successfully preserving appropriate privacy and confidentiality rights, allowing researchers to construct frameworks for analyzing and visualizing the world's population in time and space to understand agents of change, to assess their implications for society and the environment, and to develop policies and plans to meet future challenges at local, regional, national, and global scales.

For additional information, see: https://international.ipums.org/international/

GOAL 6: ENABLE COMMUNITIES OF PRACTICE

Findings: Scientific data exist in many different types and formats subject to varying legal, cultural, protection, and practical constraints. They are often used in different ways according to their contexts and have varying life cycle requirements. Data authors, managers, and users often come from different disciplinary, professional, cultural, and other settings with different needs, expectations, responsibilities, authorities, and expertise. These experts are subject to varying legal, physical, scientific, cultural, and other constraints. This diversity in data, individuals, institutions, disciplines, contexts, and cultures is a strength of the American scientific research and education system. One-size-fits-all solutions must be avoided. Solutions should support communities of practice and leverage their capabilities while promoting data integration and interoperability. Because these communities of practice are changing the way data are used and reused and the way science in these communities is done, these community processes present an opportunity for research in the social, behavioral, and other sciences.

Data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice.[19] Solutions should support such a distributed system, recognizing the diverse interests of all of the stakeholders while promoting federation and interoperability.

Diversity is a strength of the American scientific system.

Federal departments and agencies and the proposed NSTC Subcommittee will need to engage communities of practice and the leadership of community-based collections and repositories in pursuing digital data preservation and access goals. Implementing the recommendations of this report will require extensive community consultation mechanisms and a fully participatory approach to data activities. Such mechanisms should be used in promoting interoperability and federation, developing standards and formats, implementing agency requirements for deposition and access, designing capabilities and features for tools and services, and other data activities. Resulting policy and implementation plans should reflect the needs, capabilities, and interests of the broad diversity of stakeholders.

19 The role of community-based collections in a data collections universe is addressed by the National Science Board in its report "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" (see Appendix D for reference). In this report, "community-proxy" is defined as the explicit or implicit authority from the community to make choices on its behalf on issues such as collection curation, access policies, standards and ontology development, annotation content, etc.

RECOMMENDATIONS: KEY ELEMENTS

The three recommendations of this report – creation of an NSTC Subcommittee, development of agency data policies, and provisions for data management plans – are designed to work together in reshaping the digital scientific data landscape. They provide a national management framework for meeting the six goals outlined above. The key elements of each recommendation that ensure they can work in combination are set out below.

RECOMMENDATION 1 – CREATION OF AN NSTC SUBCOMMITTEE

The Subcommittee will focus on goals that are best addressed through broad cooperation and coordination, while the agencies will pursue goals specific to their respective missions and communities of practice. Examples of focus areas for the Subcommittee include the following. The priorities for these areas will be set by the members of the Subcommittee.

Extended National Coordination. Engage with federal agencies outside the NSTC Subcommittee, government at the state and local levels, other interagency coordination groups (including other relevant NSTC groups), and the commercial, academic, educational, and non-profit sectors. Goals include identifying shared opportunities and challenges, gaps and unmet needs, synergies and partnerships, and economies of scale or shared investments, which would allow the federal government to serve as both leader and partner for digital scientific data preservation and access.

International Coordination. Engage with foreign national and international agencies and entities in the government, commercial, academic, educational, and non-profit sectors. Goals include identifying shared opportunities and challenges, gaps and unmet needs, synergies and partnerships, and economies of scale or shared investments, which would allow the federal government to serve as both leader and partner for digital scientific data preservation and access.

Education and Workforce. Enable the current generation and develop the next generation of leaders and innovators in data science and technology by coordinating the activities of the NSTC Subcommittee and its partners and engaging stakeholders in other sectors.

Data Innovation Research. Coordinate research to support digital scientific data innovation. Examples of digital data innovation research include methods for assessing or achieving scalability, systems integration, and design robustness, including fault tolerance in evaluating the application of one or more inventions to particular applications or needs.

Data Systems Implementation and Deployment. Promote greater capability and capacity in implementation design and deployment of data software, hardware, and systems. The Subcommittee will encourage adoption and implementation of data preservation and access strategies, concepts, and best practices. It will also promote efficient re-use and adoption of tools and technologies to facilitate integration and interoperability.

Data Discovery and Dissemination. Promote enhanced capabilities for finding, understanding, visualizing, and interacting with data. The Subcommittee will support diverse uses through a coordinated set of relevant technologies and will disseminate information about available data.

Data Protection. Develop strategies, concepts, and tools for protecting data security, privacy, confidentiality, and intellectual property rights, and for enabling effective user access, authentication, authorization, and accounting protocols and frameworks.

Data Quality and Disposition. Develop concepts, strategies, and tools for data quality assessment and control, validation, authentication, provenance, and attribution. The Subcommittee will promote the development and sharing of best practices for disposition decision-making (i.e., which data should be kept, for how long, and by what entities), including strategies and practices for understanding the relationship between cost and benefits.[20]

Integration and Interoperability. Promote strategies, approaches, investments, and partnerships that enable the effective integration and interoperability of data and data tools, systems, services, and resources. The Subcommittee will promote the identification, use, and continuing evolution of existing standards and the development of standards where needed. This ensures coherent identification of distributed data, enhances coordination of the activities of the NSTC Subcommittee and its partners, and engages other sectors with the goal of enabling the creative use of digital scientific data in innovative combinations for purposes of discovery, innovation, and progress.

The activities of the Subcommittee should include close cooperation with the other relevant NSTC entities. Two of these are especially relevant in the digital scientific data landscape, and their relationship to the proposed new Subcommittee can be summarized as follows:

Relationship to the NITRD Subcommittee. The Networking and Information Technology Research and Development Subcommittee (NITRD) focuses on the invention phase (i.e., basic research to prototype/proof of concept) of the invention-innovation-implementation-design-deployment[21] cycle of technology change. The proposed NSTC Subcommittee on Digital Scientific Data focuses on innovation through deployment. There is necessarily both overlap and dependency between phases in this cycle, and close communication and coordination between these two groups will be implemented to manage and leverage appropriate linkage between phases. This interaction will ensure that the most promising and innovative research outputs can be considered for further development and that the research process is responsive to the real-world needs of the implementation sector.

Relationship to the Scientific Collections IWG. The Scientific Collections Interagency Working Group focuses on collections of physical objects relevant to science (e.g., biological specimens, drilling cores, fossils). Collections of digital counterparts to such physical objects (e.g., digital images or 3-dimensional digital renderings) fall within the purview of the proposed NSTC Subcommittee on Digital Scientific Data. These two groups will closely coordinate to manage the relationship between the physical and digital collection realms and to enable rational, cost-efficient decision making about digitization for preservation and access.

RECOMMENDATION 2 - AGENCY DIGITAL DATA POLICY

The second key element of the strategic framework is that appropriate departments and agencies lay the foundations for agency digital scientific data policy and make the policy publicly available. In laying these foundations, agencies should consider all components of a comprehensive policy to address the full data management life cycle. Examples of such components include the following:

20 The benefits of digital preservation must be continuously weighed against the costs. Assessment of benefits must rely extensively on input from the relevant stakeholder communities, be conducted openly, and be consistent with the mission of the relevant department or agency. Such assessment should include consideration of the full range of benefits, both tangible and intangible. The assessment should compare the costs of preserving a data set with the possibility and costs of regenerating the data. When reproducing data is not possible, preservation should be the preferred choice where feasible. Cost analyses should be informed by comprehensive and reliable information. Similar analyses should be conducted for plans to digitize physical artifacts (books, documents, reference samples and specimens, etc.) for preservation and access. Recognizing that current analyses are limited by the lack of comprehensive economic theory and management frameworks for long-term digital preservation, agencies should work together to support research and development to improve the conceptual foundations and methodologies in this area.

21 We distinguish between "invention" and "innovation" in the manner of Schumpeter (see Schumpeter, JA. *Business Cycles*. New York. McGraw Hill. 1939), with "invention" referring to the discovery of new concepts or devices and "innovation" as the creative use, modification, or combination of existing concepts and devices for desired applications.

Statement of guiding principles for digital scientific data preservation and access. The principles should provide clear guidelines for those conducting the data planning and implementation activities of the agency and for those seeking to partner with the agency in pursuing shared data goals. This includes criteria for determining whether data are appropriate for preservation and access. Further, the principles must be in accordance with the provisions of the *Paperwork Reduction Act* (44 U.S.C. 3501 et seq.), *OMB Circular A-130*, the *America COMPETES Act*, the *Data Quality Act*, the *Federal Funding Accountability and Transparency Act* (FFATA), and other applicable policy, regulatory, and statutory requirements. The agency digital data policy should cite the relevant governing documents wherever appropriate.

Assignment of responsibilities. The roles of agency offices and officials in implementing the agency digital data policy should be described to ensure clear lines of authority and accountability and to provide transparency for those working within and outside the agency on digital data matters. This should include provisions for a designated, cognizant senior science official serving as Science Data Officer to coordinate the digital data activities of the agency and to serve as representative to the Subcommittee on Digital Scientific Data.

Description of mechanisms for access to specialized data policies. Agencies may support various communities of practice and distinct data types, formats, and contexts, and they may have differing programmatic goals, needs, and resources. Such agencies should have a harmonized suite of corresponding, specialized data policies. The comprehensive agency digital data policy should describe mechanisms to provide easy and transparent access to the agency's full portfolio of specialized data policies.

Statement of intentions and mechanisms for cooperation, coordination, and partnerships. The agency digital data policy should describe the agency's intentions and mechanisms for cooperation, coordination, and partnerships across sectors. Such sectors can include government at the national, state, or local levels, as well as industry, academia, education, non-profits, and international entities.

Provisions for updating and revisions. The agency digital data policy must be a living document if it is to remain relevant and effective in a dynamic landscape. The policy should describe the mechanisms to be used for updating and revising the document to ensure it is responsive to change and opportunity.

RECOMMENDATION 3 - DATA MANAGEMENT PLAN ELEMENTS

The third key element of the strategic framework is for all agencies to promote a data management planning process for projects that generate preservation data. This includes preparing a data management plan in proposals for activities that will generate digital scientific data. Examples of elements that should be considered in such a data management plan are listed below. This listing can be consulted by agencies in developing an appropriate portfolio of specialized data management policies, with each policy crafted for the community and context in which a particular project or projects will be conducted. Each specialized policy may include or omit any of the elements listed below or add others as appropriate to the particular application or context.

Description. Brief, high-level description of the digital scientific data to be produced.

Impact. Discussion of possible impact of the data within the immediate field, in other fields, and any broader, societal impact. Indicate how the data management plan will maximize the value of the data.

Content and Format. Statement of plans for data and metadata content and format, including description of documentation plans and rationale for selection of appropriate standards. Existing, accepted standards should be used where possible. Where standards are missing or inadequate, alternate strategies for enabling

data re-use and re-purposing should be described, and agencies should be alerted to needs for standards development or evolution.

Protection. Statement of plans, where appropriate and necessary, for protection of privacy, confidentiality, security, intellectual property and other rights.

Access. Description of plans for providing access to data. This should include a description and rationale for any restrictions on who may access the data under what conditions and a timeline for providing access. This should also include a description of the resources and capabilities (equipment, connections, systems, expertise, etc.) needed to meet anticipated requests. These resources and capabilities should be appropriate for the projected usage, addressing any special requirements such as those associated with streaming video or audio, movement of massive data sets, etc.

Preservation. Description of plans for preserving data in accessible form. Plans should include a timeline proposing how long the data are to be preserved, outlining any changes in access anticipated during the preservation timeline, and documenting the resources and capabilities (e.g., equipment, connections, systems, expertise) needed to meet the preservation goals. Where data will be preserved beyond the duration of direct project funding, a description of other funding sources or institutional commitments necessary to achieve the long-term preservation and access goals should be provided.

Transfer of Responsibility. Description of plans for changes in preservation and access responsibility. Where responsibility for continuing documentation, annotation, curation, access, and preservation (or its counterparts, de-accessioning or disposal) will move from one entity or institution to another during the anticipated data life cycle, plans for managing the exchange and documentation of the necessary commitments and agreements should be provided.

Appendix A

Interagency Working Group on Digital Data
Terms of Reference (Charter)

A. INTERAGENCY WORKING GROUP ON DIGITAL DATA TERMS OF REFERENCE (CHARTER)

TERMS OF REFERENCE of the
INTERAGENCY WORKING GROUP ON DIGITAL DATA
COMMITTEE ON SCIENCE
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

PREAMBLE

The Interagency Working Group on Digital Data (the "Interagency Working Group" or IWG") is hereby established by the Committee on Science ("the Committee" or "COS"). The IWG serves as a part of the internal deliberative process of the Committee on Science.

PURPOSE

The purpose of the IWG is to develop and promote the implementation of a strategic plan for the federal government to cultivate an open interoperable framework to ensure reliable preservation and effective access to digital data for research, development, and education in science, technology, and engineering. For the purposes of this document, digital data are defined as any information that can be stored digitally and accessed electronically, with a focus specifically on data used by the federal government to address national needs or derived from research and development funded by the federal government. Analog data digitized for storage are also included. The term "agencies" refers to federal departments, agencies, directorates, institutes, and other organizational entities. While emphasis is on U.S. federal entities, scientific data management crosses national boundaries, and the work of this IWG will take into account international dimensions of a data framework.

The IWG will provide a means for coordinating policy, programs, and budgets among federal agencies and with partners in other sectors. This includes identifying and integrating requirements, conducting joint program planning, and developing joint strategies for digital data preservation and access activities conducted by agency members of the IWG. The strategic plan should provide for cost-effective cooperation and coordination among agencies and with the science, technology, and engineering research and development communities, and with international partners and counterparts, as appropriate, to identify best practices, to encourage shared solutions to key challenges, and to implement coordinated strategies and policies for managing digital data.

SCOPE

The scope of activities for the IWG includes:

- *Developing a strategic plan for the federal government, working in partnership with other sectors, to enable reliable preservation of and effective access to digital data, appropriately protected, in science, technology, and engineering;*
- *Promoting the implementation of the strategic plan through coordination among federal agencies and through partnerships with other sectors;*
- *Developing strategic requirements for an open interoperable data framework;*
- *Promoting communications among developers and users of digital data for research, development, and education in science, technology, and engineering, to help ensure that their digital data needs are addressed;*
- *Assuring necessary international collaboration, access, and interoperability; and*
- *Ensuring that the activities of the IWG are informed by and not duplicative of the ongoing activities of other groups in areas such as electronic health care and medical records.*

FUNCTIONS

The IWG has the following functions and activities:

- *Facilitating interagency digital data strategic plan development and implementation, including:*
  - *Assessing the current status of digital data generation, archiving, preservation, and access among federal agencies;*
  - *Providing a forum for agencies to exchange program-level information about agency digital data activities;*
  - *Recognizing agency priorities and identifying interagency priorities in digital data, identifying any gaps in the federal strategy related to those areas, and promoting interagency coordination to address these gaps;*
  - *Identifying opportunities for domestic and international collaboration, coordination, and leveraging among agencies in specific digital data areas;*
  - *Coordinating policy, programs, and budgets for implementing the strategic plan.*
- *Facilitating interoperability broadly and recommending means and processes to achieve it, including mechanisms such as standards evolution and development;*
- *Facilitating coordination and cooperation with the research, development, and education communities;*
- *Facilitating a strong interagency planning effort;*
- *Maintaining and overseeing coordinating groups in specific science or technology areas;*
- *Maintaining active awareness of data sets in technical areas other than science, technology, and engineering, and within the international community;*
- *Submitting an annual progress report to the Committee.*

MEMBERSHIP

The following federal agencies are represented on the IWG:

- *Department of Agriculture*
- *Department of Commerce*
- *Department of Defense*
- *Department of Education*
- *Department of Energy*
- *Department of Health and Human Services*
- *Department of Homeland Security*
- *Department of the Interior*
- *Department of Labor*
- *Department of Justice*
- *Department of State*
- *Department of Transportation*
- *Department of the Treasury*

- *Department of Veterans Affairs*
- *Central Intelligence Agency*
- *Environmental Protection Agency*
- *Library of Congress*
- *National Aeronautics and Space Administration*
- *National Archives and Records Administration*
- *National Science Foundation*
- *The Smithsonian Institution*
- *US Army Corps of Engineers*

The following councils and offices shall participate in IWG activities:

- *Council on Environmental Quality*
- *Domestic Policy Council*
- *Homeland Security Council*
- *National Economic Council*
- *National Security Council*
- *Office of Management and Budget*
- *Office of Science and Technology Policy*

LEADERSHIP AND OPERATIONS

Co-Chairs of the IWG shall be named by the Co-Chairs of the Committee. Intra- and inter-agency coordination, fact finding, coordinating group efforts, and planning shall occur during and/or between the formal, scheduled IWG meetings.

INTERACTIONS WITH OTHER ORGANIZATIONS

The IWG may interact with other government organizations including the NSTC Committee on Technology (COT), the Networking and Information Technology R&D (NITRD) Subcommittee, which reports to the COT, and the NITRD National Coordination Office. The IWG may also interact with federal advisory bodies such as the President's Council of Advisors on Science and Technology (PCAST). The IWG may interact with and receive ad hoc advice from other interagency groups such as CENDI and the Federal CIO Council, and from private sector groups, professional societies, and other non-government organizations such as the National Academies of Science and Engineering, the Institute of Medicine, and the National Research Council as consistent with the *Federal Advisory Committee Act.*

TERMINATION

Unless renewed by the Co-Chairs of the Committee on Science prior to its expiration, the IWG shall terminate no later than March 31, 2009.

DETERMINATION

We hereby determine that the formation of this Interagency Working Group is in the public interest in connection with the performance of duties imposed on the Executive Branch by law, and that such duties can best be performed by such a group.

Appendix B

Digital Data Life Cycle

B. The Digital Data Life Cycle

Exhibit B-1. Digital Data Life Cycle Model

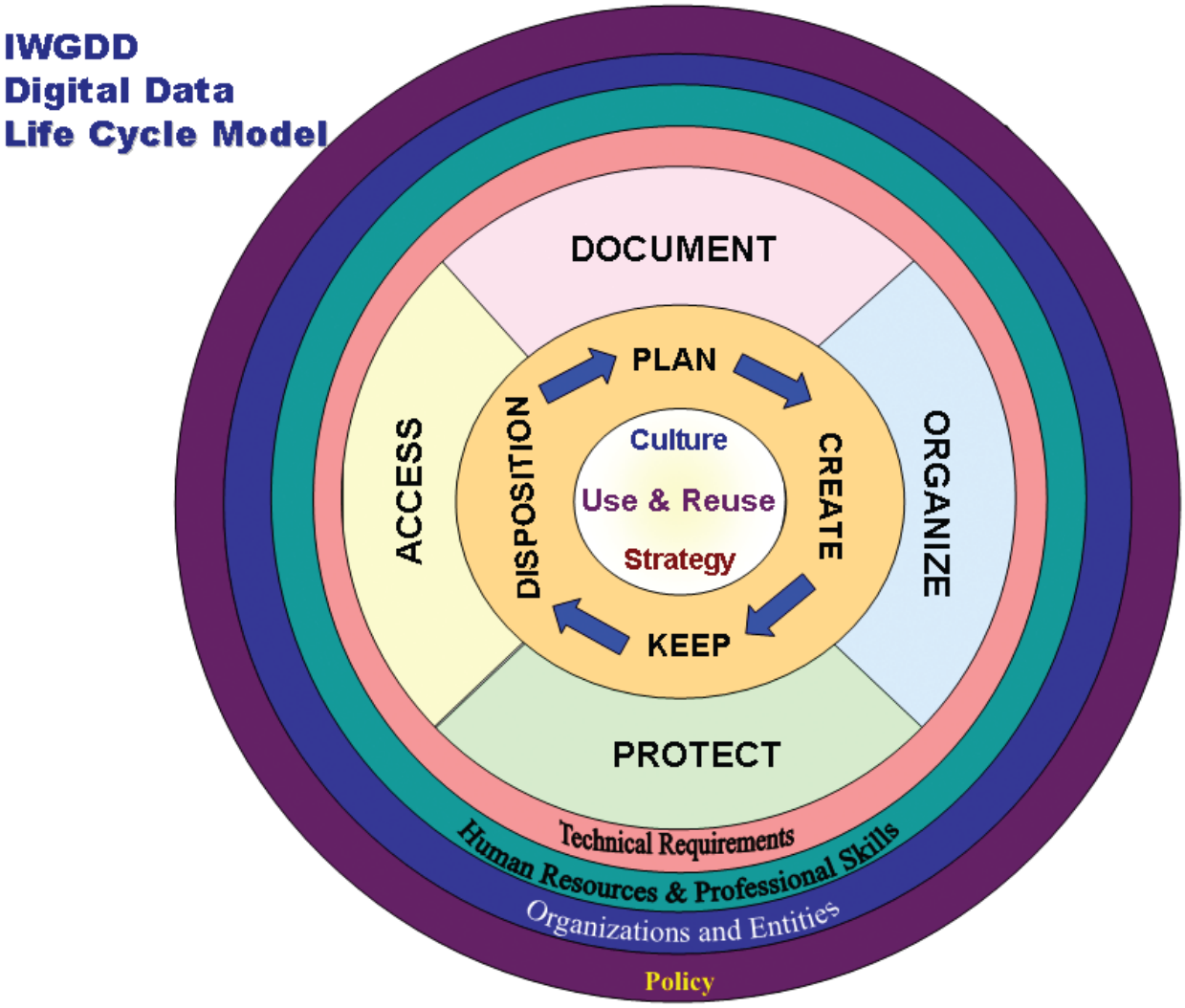


Exhibit B-2. Life Cycle Functions for Digital Data*

- *Plan*
  - *Determine what data need to be created or collected to support a research agenda or a mission function*
    - Identify and evaluate existing sources of needed data
  - *Identify standards for data and metadata format and quality*
  - *Specify actions and responsibilities for managing the data over their life cycle*
- *Create*
  - *Produce or acquire data for intended purposes*
  - *Deposit data where they will be kept, managed and accessed for as long as needed to support their intended purpose*
  - *Produce derived products in support of intended purposes; e.g., data summaries, data aggregations, reports, publications*
- *Keep*
  - *Organize and store data to support intended purposes*
    - Integrate updates and additions into existing collections
    - Ensure the data survive intact for as long as needed
- *Acquire and implement technology*
  - *Refresh technology to overcome obsolescence and to improve performance*
  - *Expand storage and processing capacity as needed*
  - *Implement new technologies to support evolving needs for ingesting, processing, analysis, searching and accessing data*
- *Disposition*
  - *Exit Strategy: plan for transferring data to another entity should the current repository no longer be able to keep it*
  - *Once intended purposes are satisfied, determine whether to destroy data or transfer to another organization suited to addressing other needs or opportunities*

**Life cycle functions are necessarily sequential in any research or other program, but the same body of data may go through multiple cycles as it is used by different entities or for different purposes.*

Exhibit B-3. Data Management Functions for Scientific and Technical Data

{These functions occur across all phases of the data life cycle}

- *Document*
  - *Define standards for data content, form, metadata, quality, frequency of updates, etc.*
  - *Create/maintain metadata*
  - *Document data history: provenance and lineage, actual data collection and processing (e.g., calibration, geo-referencing, noise reduction)*
  - *Note anomalies and lacunae*
  - *Record disposition decisions and actions*

- *Organize*
  - *Design and implement data architecture, engineering and structures*
  - *Conform to standards*

- *Protect*
  - *Implement quality control*
    - Verify and validate data on ingest
    - Ensure integrity and validity of any transformations or derived products
  - *Implement access restrictions*
    - Respect property rights
    - Protect privacy and confidentiality
  - *Guarantee availability to authorized users*
    - Define user roles and privileges
    - Qualify individual users
  - *Guarantee trustworthiness and authenticity*
    - Function as a trusted repository
    - Implement, maintain and monitor the security of system and the assets stored in it
    - Implement methods for ensuring and verifying authenticity

- *Access*
  - *Acquire data from existing sources*
  - *Catalogue and describe as to content, quality, availability, etc.*
  - *Ensure coherent identification of distributed data*
  - *Disseminate information about available data*
  - *Support diverse uses through an appropriate variety of technologies*
  - *Support a variety of methods of discovery, analysis, repurposing, dissemination, presentation*

Appendix C

Organizations, Individuals, Roles, Sectors, and Types

C. Organizations, Individuals, Roles, Sectors, and Types

1. Entities by Role
2. Entities by Individual
3. Entities by Sector
4. Individuals by Role
5. Individuals by Life Cycle Phase/Function
6. Entities by Life Cycle Phase/Function

Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
Research Projects	Collect or produce data through original research Develop and validate improved testing methods Develop data collection or production instruments, techniques, or processes Operate laboratories or observatories Preserve original and/or derived data Produce publications from research Produce refined data products through calibration, geo-referencing, or other enhancement of raw data Collect data from other producers Provide access to bibliographic data about research Provide access to original or derived data	European Bioinformatics Institute American National Election Survey Framingham Heart Study General Social Survey National Toxicology Program NIST Physics Laboratory Panel Study of Income Dynamics UNAVCO
Data Centers /Statistical Agencies	Collect data from other producers Collect or produce data through original research Combine data from multiple sources Develop data collection or production instruments, techniques, or processes Preserve original or derived data sets Promote collaboration on production, dissemination or management of data Provide access to original or derived data Provide resources or services for analyzing or processing data Publish research results	Government Agencies: Science Data Centers Center for Earth Resources Observation and Science National Climactic Data Center National Oceanographic Data Center NSF's Census Research Data Centers
	Collect or produce data through original research Combine data from multiple sources Collect data from other producers Provide resources or services for analyzing or processing data Provide access to original or derived data Provide financing for projects in other organizations to produce, disseminate, or access data Provide training on information dissemination and access	Government Agencies: Statistical Agencies Bureau of Census Census State Data Center Program Division of Science Resources Statistics NSF Economic Research Service
	Analyze and revise data to improve their quality Collect data from other producers Collect or produce data through original research Combine data from multiple sources Develop data collection or production instruments, techniques, or processes Operate laboratories or observatories Preserve original or derived data sets Promote collaboration on production, dissemination or management of data Provide access to bibliographic or other reference data Provide access to original or derived data Provide resources or services for analyzing or processing data Provide training on data analysis, processing or management	Private Sector Centers/Activities Chandra X-ray Center at the Smithsonian Astrophysical Observatory Economic and Social Data Service UK National Optical Astronomy Observatory Space Telescope Science Institute Worldwide Protein Data Bank

Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
Libraries	Analyze and revise data to improve their quality or usefulness Collect derived data products, principally publications Combine data from multiple sources Convert analog information or materials to digital formats Create bibliographic and other reference data Develop instruments, techniques, or processes for data collection or production, processing, management, or dissemination Develop and enhance software tools that will enable gene discovery Preserve publications Preserve original and/or derived data Provide access to bibliographic or other reference data Provide access to publications Provide financing for projects in other organizations to produce, disseminate, or access data	National Library of Medicine Wellcome Library
Information Service Providers	Collect or produce data through original research Conduct data management research Promote improved data management Provide data management services, tools or facilities Provide tools for data dissemination Promote collaboration on production, dissemination or management of data Promote data sharing Collect data from other producers Publish research results Provide access to bibliographic or other reference data Provide access to publications Provide access to original or derived data Preserve original or derived data sets Provide training materials for data analysis Provide training on use of scientific data in different contexts Research and develop computational capabilities for science and engineering	Astrophysics Data System Inter-university Consortium for Political and Social Research Journal of the American Statistical Association Data Archive National Association of Health Data Organizations National Fusion Grid Semantic Web for Health Care and Life Sciences Interest Group Sociometrics Social Science Electronic Data Library
Archives	Articulate criteria and tools for assessing compliance with standards Collect data from other producers Develop and promulgate data standards Preserve original or derived data sets Preserve publications Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications Provide resources or services for analyzing or processing data Provide training on data analysis, processing or management Provide training on life cycle management	National Archives and Records Administration National Data Archive on Child Abuse and Neglect Open Archives Initiative UK Data Archive
Museums	Collect or produce data through original research Convert analog information or materials to digital formats Operate laboratories or observatories Preserve original or derived data sets Provide access to bibliographic or other reference data Provide access to original or derived data Provide resources or services for analyzing or processing data Publish research results	Field Museum Muséum national d'Histoire naturelle Smithsonian Museums Yale Peabody Museum

Exhibit C-1. Entities by Role

ENTITY TYPE	ROLE	EXAMPLES
National / International Infrastructure	Develop and promulgate data standards Organize/sponsor conferences Promote collaboration on production, dissemination or management of data Promote data sharing Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications	Council of European Social Science Data Archives Global Price and Income History Group – University California/Davis Integrated Public Use Microdata Series International Luxembourg Income Study National Biological Information Infrastructure National Spatial Data Infrastructure
STI Centers	Collect data from other producers Determine policies regarding collection, content, quality, peer review and dissemination of data Promote collaboration on production, dissemination or management of data Provide access to bibliographic or other reference data Provide access to original or derived data Provide access to publications Provide data management services, tools or facilities	DoD, Defense Technical Information Center DOE, Office of Scientific and Technical Information NASA Technical Reports Server
Computer Centers	Enable formation of virtual organizations through computational and data grids Preserve original or derived data Provide facilities and vehicles for collaboration Provide tools for data processing, access, and use Provide training on use of tools for data processing, access, and use Research and develop computational capabilities for science and engineering Store and process data	National Center for Supercomputing Applications Renaissance Computing Institute San Diego Supercomputer Center
Standards Bodies	Develop and promulgate data standards Promote data sharing Provide training on implementation of data standards Publish books and periodicals on data standards and their use Register service providers deemed competent in data standards	Clinical Data Interchange Standards Consortium Consultative Committee on Space Data Systems
Audit/ Accreditation Bodies	Accredit laboratories' technical qualifications and competence to carry out specific calibrations or tests Articulate criteria and tools for assessing compliance with standards Audit data production, management, preservation and dissemination activities	Government Accountability Office NIST, National Voluntary Laboratory Accreditation Program Research Libraries Group
Information Distributors	Collect data from other producers Provide access to bibliographic or other reference data Provide peer review of publications Provide access to publications Provide access to original or derived data Publish research results Provide training on information dissemination and access Preserve original or derived data Preserve publications	EconData.Net Elsevier International Network for the Availability of Scientific Publications Internet Scientific Publications *Journal of the American Medical Association* Thomson Reuters
Hardware Software Developers/ Suppliers	Provide tools for data production, processing, preservation, access, and use Research and develop computational capabilities for science and engineering	IT industry Open source software collaborations

Exhibit C-2. Entities by Individuals

ENTITIES	Data Center Scientists	Data Scientists	Librarians	Archivists	Record Managers	Researchers	Modelers	Students	Information & Data Management Specialists	Computer Scientists, Engineers, & IT Specialists	Journalists, Science Writers	Research Program Directors/Policy Makers
Research Projects						X	X	X	X	X		
Data Centers /Statistical Agencies	X	X			X	X	X	X	X	X		
Libraries			X		X				X	X		X
Information Service Providers (e.g., Catalog Services)	X				X				X	X		
Archives				X	X				X			
Museums		X			X				X			
National/International Infrastructure (e.g., NBII, NSDI)									X	X		X
STI Centers (OSTI, CASI, DTIC)		X	X		X		X	X	X	X		
Computer Centers (SDSC, NCSA)						X	X		X	X		
Standards Bodies (CCSDS)									X			
Audit/Accreditation Bodies									X			
Information Distributors (Including Publishers, Conference Organizers, Press)									X	X	X	
Hardware/Software Developers/Suppliers									X	X		

Exhibit C-3. Entities by Sector

ENTITIES	Government			Education			Multi-Sector Collaboration		Research & Development Institutions	Not for Profit / NGO	For-Profit
	Legislative	Executive	Judicial	K-12	Vocational	Academic Higher Education	International	National			
Research Projects		X		X	X	X	X	X	X	X	X
Data Centers /Statistical Agencies		X				X	X	X	X	X	X
Libraries	X	X		X	X	X			X		
Information Service Providers		X				X	X	X	X	X	X
Archives		X				X			X	X	
Museums		X				X				X	
National/International Infrastructure		X [1]					X	X	X		
STI Centers		X									
Computer Centers						X			X	X	X
Standards Bodies		X [2]					X [3]	X		X [4]	
Audit/Accreditation Bodies		X									
Information Distributors						X			X	X	X
Hardware and Software Developers/Suppliers		X [5]					X [6]	X		X	X

1 Some agencies provide leadership for multi-sector collaboration.
2 Standards bodies across the sector include NIST.
3 Standards bodies across the sector include ISO.
4 Standards bodies across the sector include OGC.
5 Agencies may develop and distribute software tools/models, etc.
6 Collaboratives may develop software tools.

Exhibit C-4. Individuals by Role

INDIVIDUALS	ROLE
Data center scientists	Disciplinary scientists who work in data centers and develop a special expertise in data management and data science. Plan, manage, give scientific oversight and input to all phases of the data management cycle within the data center/archive, including scientific specifications for software development to support data processing and archival operations; management of these operations; validation and verification of data products; interoperability links with other archives and the literature; and design and management of data re-use products such as data mining from archival products and catalog creation.
Data scientists	Scientists who come from information or computer science backgrounds but learn a subject area and may become scientific data curators in disciplines and advance the art of data science. Focus on all parts of the data life cycle.
Librarians	Focus on the functions — keeping and disposing of information and planning in regard to it. Generally not in the "creation" role. Collect relevant information to manage through the life cycle based on scope and cover of the library mission. Usually collect information from a variety of creating entities. Focus on the access and use of data. Individuals work to standards of the library profession in organizing, protecting, accessing, and documenting data in the two main functions of the life cycle.
Archivists	Select, preserve, and provide access to data and related information as records (i.e., collections organically produced, structured, and interrelated in the course of scientific activity). Preserve the original form, content, and structure and sufficient contextual information about the producers and the activities in which the data were produced to enable correct interpretation and informed judgment on their reliability and limitations. Not in the creation role.
Record Managers	Play a functional role between the creators of information and the archivists from a particular institutional perspective. Focus on keeping and disposing, with emphasis on protecting and documenting for institutional use.
Researchers	Conceive, plan, experiment and analyze data to produce results for scientific publication. Modelers who use data (their own or that of other scientists) to develop and run models can be considered a specific class of researcher. While most individual researchers focus primarily on data collection and analysis and do not usually focus on documentation or preservation, some may carry out the full life cycle function for the data they create.
Students	Assist researchers and or participate in experiment for school/thesis work. May be data producers.
Information and Data Management Specialists	Provide operational support to data management operations, including pipeline data processing, ingest into the archive, archive management, data access and distribution oversight, production of use statistics, etc. Play the roles similar to librarians, archivists, or records managers, but do not necessarily work to the library profession standards.
Computer Scientists, Engineers, and IT Specialists	Design and develop software to support data management operations (processing, archiving, distribution, etc.) following scientist's specifications. Design and develop (acquire) computer systems to support these operations, ensuring speed, security, etc., as required by the project. This includes acquiring hardware, setting up networks, and acquiring and installing systems software.
Journalists, Science Writers	Translate data from highly scientific fields to be available to other audiences with various levels of scientific understanding.
Research Program Directors/Policy Makers	Provide overall strategic direction and resource allocation for research programs. Focus on the planning functions for data.

Exhibit C-5. Individuals by Life Cycle Phase/Function

INDIVIDUAL	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Center Scientists	X	X	X	X	X	X	X	X
Data Scientists	X	X	X	X	X	X	X	X
Librarians	X		X	X	X	X	X	X
Archivists	X		X	X	X	X	X	X
Record Managers			X	X		X		X
Researchers	X	X			X			
Students	X	X			X			
Information and Data Management Specialists		X	X	X	X	X	X	X
Computer Scientists, Engineers, and IT Specialists	X	X	X					
Journalists, Science Writers	X	X	X	X	X	X	X	X
Research Program Directors/Policy Makers	X							

Exhibit C-6. Entities by Life Cycle Phase/Function

ENTITIES	Data Life Cycle Phase				Data Management Functions			
	Plan	Create	Keep	Dispose	Access	Document	Organize	Protect
Data Projects	X	X	X	X	X	X	X	X
Data Centers / Statistical Agencies	X	X	X	X	X	X	X	X
Libraries			X	X	X	X	X	X
Information Service Providers	X	X	X	X	X	X	X	X
Archives			X	X	X	X	X	X
Museums			X	X	X	X	X	X
National/International Infrastructure					X	X	X	X
STI Centers					X	X	X	X
Computer Centers					X	X	X	X
Standards Bodies						X	X	
Audit/Accreditation Bodies						X	X	
Information Distributors		X	X	X	X	X	X	X
Hardware Software Developers/Suppliers					X	X	X	X

Appendix D

Related Documents

A. Related Documents

IWGDD Key Digital Data Bibliographical References (Revised 04/08/08)

"A Fresh Look at the Reliability of Long-term Digital Storage." Baker, Roussopoulos, Shah, Maniatis, Bungale, Rosenthal, and Giuli. White Paper. March 2006. http://lockss.org/locksswiki/files/Eurosys2006.pdf

"A Strategy for the National Data Spatial Data Infrastructure." Federal Geographic Data Committee, 1997. http://www.fgdc.gov/nsdi/policyandplanning/nsdi-strategic-plans

"Audit of NSF's Policies on Public Access to the Results of NSF-Funded Research." National Science Foundation, Office of Inspector General. February 2006. OIG 06-2-004. http://www.nsf.gov/oig/06-2-004_final.pdf

"Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research." Government Accountability Office, Report to Congressional Requesters. GAO-07-1172. September 2007. http://republicans.energycommerce.house.gov/Media/File/News/10.22.07_GAO_Report_Data_Sharing_Climate_Research.pdf

"Data Management for the North America Carbon Program." Conkright, Margarita. National Aeronautics and Space Administration. January 2005.

"The Data Reference Model Version 2.0." Federal Enterprise Architecture Program. November 17, 2005. http://www.whitehouse.gov/omb/egov/documents/DRM_2_0_Final.pdf

"Dealing with Data: Roles, Rights, Responsibilities, and Relationships." Consultancy Report. Dr. Elizabeth Lyon, UKOLN, University of Bath. June 2007. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

"Department of Defense: Information Sharing Strategy." Office of the Chief Information Officer. White Paper/Strategy. May 2007. YouTube Video. http://www.youtube.com/watch?v=85OW0IyeS8s

"EIA 859 Handbook Highlights." National Archives and Records Administration. September 2004. https://acc.dau.mil/GetAttachment.aspx?id=33771&pname=file&lang=en-US&aid=6882

"Electronic Resource Preservation and Access Network Training: The Selection, Appraisal and Retention of Digital Scientific Data." Highlights of an ERPANET/CODATA Workshop. Committee on Data for Science and Technology. October 2004. http://www.jstage.jst.go.jp/article/dsj/3/0/3_227/_article

"Environmental Sampling, Analysis and Results, Data Standards Overview of Component Data Standards." Standard No.: EX 000001.0. Environmental Data Standards Council. January 2006. Standard No. EX 000007.1. January 2006.

"Federal Enterprise Architecture Records Management Profile." Version 1.0. Office of Management and Budget. December 2005. http://www.archives.gov/records-mgmt/pdf/rm-profile.pdf.

"Implementation Guide for Data Management GEIA-HB-859. " January 2006. Available for purchase at http://www.techstreet.com/cgi-bin/detail?product_id=1256205.

"Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." National Science Board, September 2005. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

"National Science Foundation Investing in America's Future." Strategic Plan FY 2006-2011, September 2006. http://www.nsf.gov/pubs/2006/nsf0648/NSF-06-48.pdf

"NSF's Cyberinfrastructure Vision for 21st Century Discovery." National Science Foundation, Cyberinfrastructure Council, September 26, 2005, Version 4.0. http://www.nsf.gov/od/oci/CI-v40.pdf

"Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources." Commission on Physical Sciences, Mathematics, and Applications, National Research Council, 1995. Available for purchase at http://www.nap.edu/catalog.php?record_id=4871.

"Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data." National Research Council Report. Committee on the Human Dimensions of Global Change, 2007. http://books.nap.edu/openbook.php?isbn=0309104149

"Records Management Guidance for Agencies Implementing Electronic Signature Technologies." National Archives and Records Administration. October 2000. http://www.archives.gov/records-mgmt/faqs/pdf/electronic-signiture-technology.pdf

"Science, Government and Information." The Weinberg Report to the President's Science Advisory Committee (PSAC), 1963.

"Scientific Data and Information: A Report of the Committee on Scientific Planning and Review Assessment Panel." International Council for Science. December 2004. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

"Soil Biodiversity Thematic Programme Data Management Plan." National Environment Research Council (NERC). June 2000. http://soilbio.nerc.ac.uk/Download/datamanplan-V2.doc

"Standard Data Management GEIA-859." Government Electronics and Information Technology Association. August 2004. Available for purchase at http://sunzi1.lib.hku.hk/ER/detail/hkul/3163032.

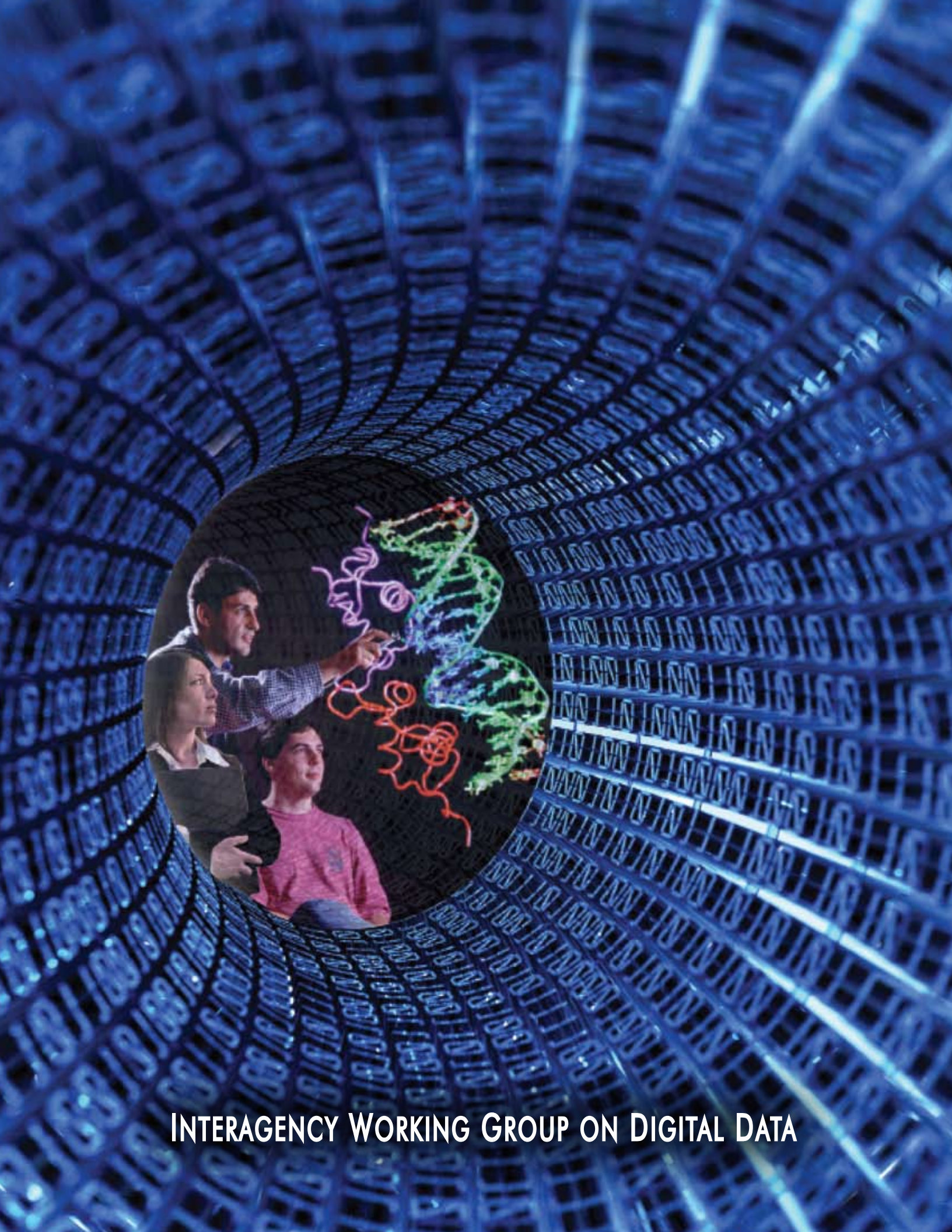
"The Facts of the Matter: Finding, Understanding, and Using Information about Our Physical World." Workshop Report on a Future Information Infrastructure for the Physical Sciences hosted by DOE and NAS, May 2000. http://www.osti.gov/physicalsciences/wkshprpt.pdf

"The Nation's Environmental Data: Treasures at Risk." National Oceanic and Atmospheric Administration. August 2001. http://www.ngdc.noaa.gov/noaa_pubs/treasures.shtml

"The Role of Scientific and Technical Data and Information in the Public Domain." Proceedings of a Symposium, National Research Council, 2003. http://www.nap.edu/catalog.php?record_id=10785

"The State of Data Management in the DOE Research and Development Complex." Report of the Meeting, "DOE Data Centers: Preparing for the Future," held July 14-15, 2004, Oak Ridge, Tennessee. November 2004. http://www.osti.gov/publications/2007/datameetingreport.pdf

"To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering." Report to National Science Foundation from Association of Research Libraries (ARL) Workshop, September 2006. http://www.arl.org/bm~doc/digdatarpt.pdf



Interagency Working Group on Digital Data